1. Dataset Partition Strategy:

A. Training Dataset (80%)

Given that there are numerous features and potential complexity in the dataset, having a larger training set will help the model learn the intricate patterns more effectively, promoting a better generalization performance. Considering the dataset contains multiple features that can interact in complex ways like seasonality effects involving 'arrival_date_month', and 'arrival_date_week_number', a larger training set can better facilitate capturing these interactions.

B. Validation Dataset (10%)

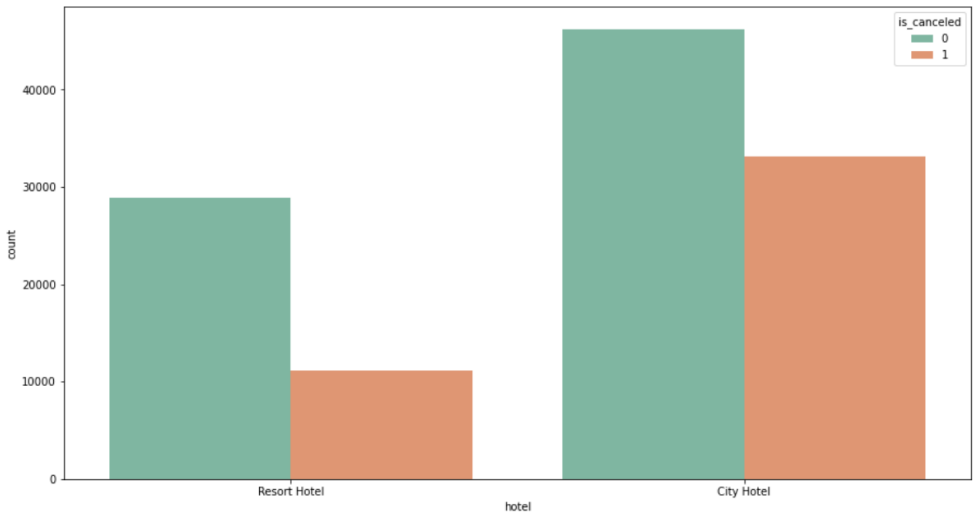
A validation set is essential for hyperparameter tuning and model selection. It helps in fine-tuning the model based on a set of data that is not seen by the model during training, helping to prevent overfitting and select the best-performing model.

C. Test Dataset (10%)

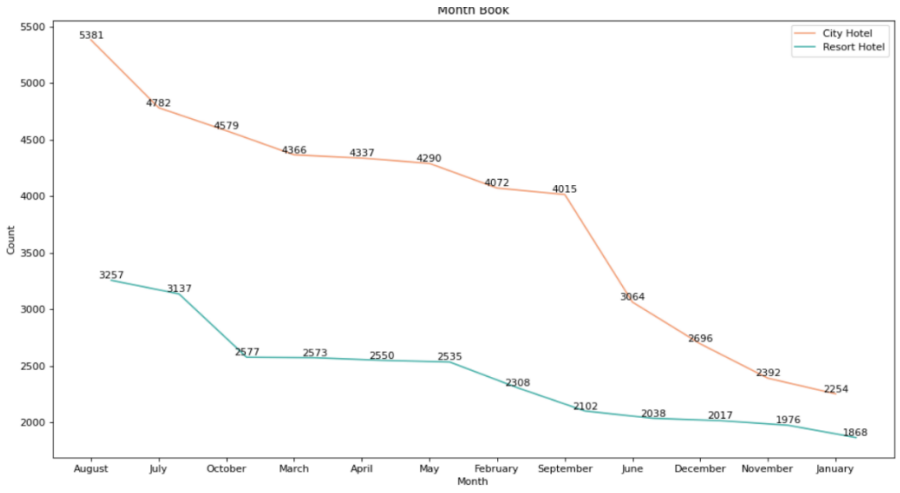
The test set is used for an unbiased evaluation of the final model fit, as it should constitute data not seen by the model during training or validation phases. It gives insights into how the model will perform in the real world, providing a final check to avoid overfitting and ensuring that the model has generalized well to new, unseen data.

2. EDA Analysis: Visualizations:

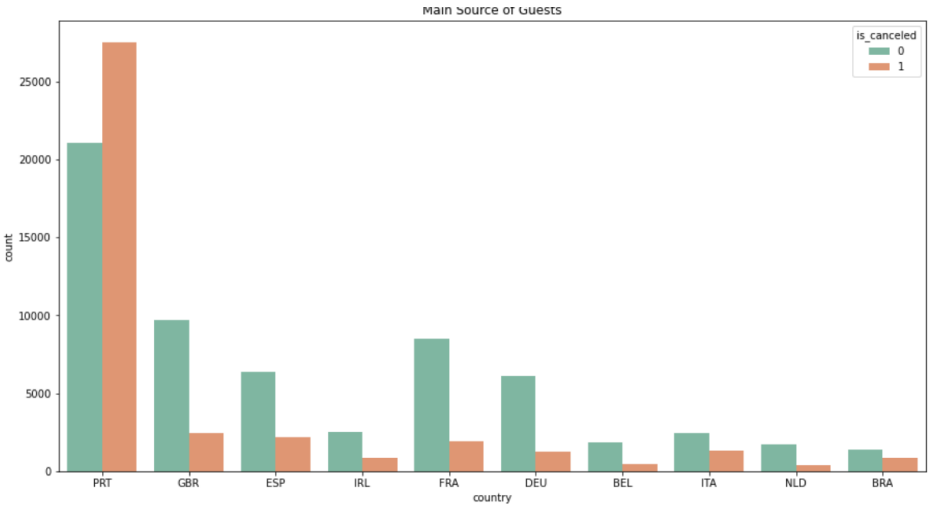
A.



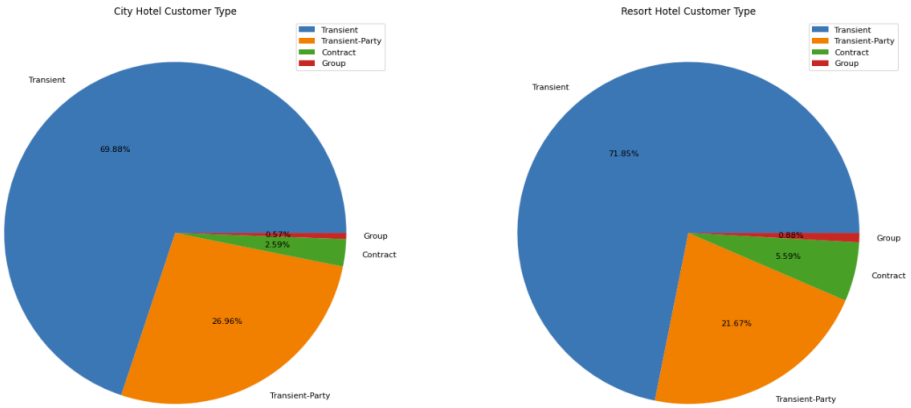
B.


Month Book
City Hotel
Resort Hotel
Count
Month
August
July
October
March
April
May
February
September
June
December
November
January
5381
4782
4579
4366
4337
4290
4072
4015
3064
2696
2392
2254
3257
3137
2577
2573
2550
2535
2308
2102
2038
2017
1976
1868

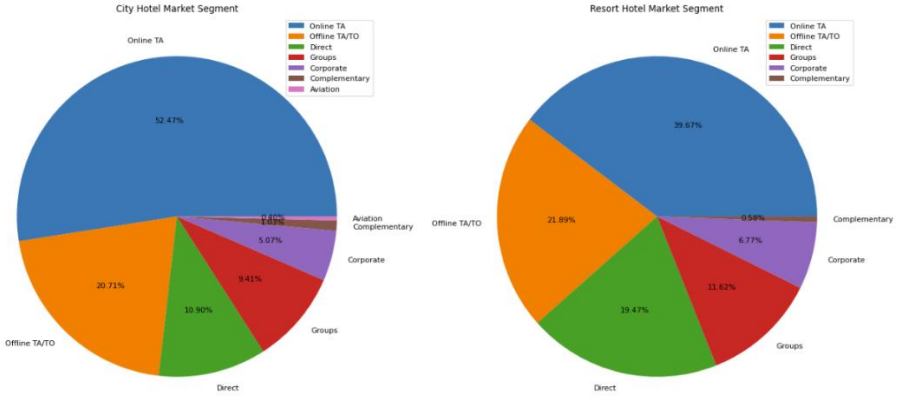
C.


Main Source of Guests
is_canceled
0
1
count
country
PRT
GBR
ESP
IRL
FRA
DEU
BEL
ITA
NLD
BRA

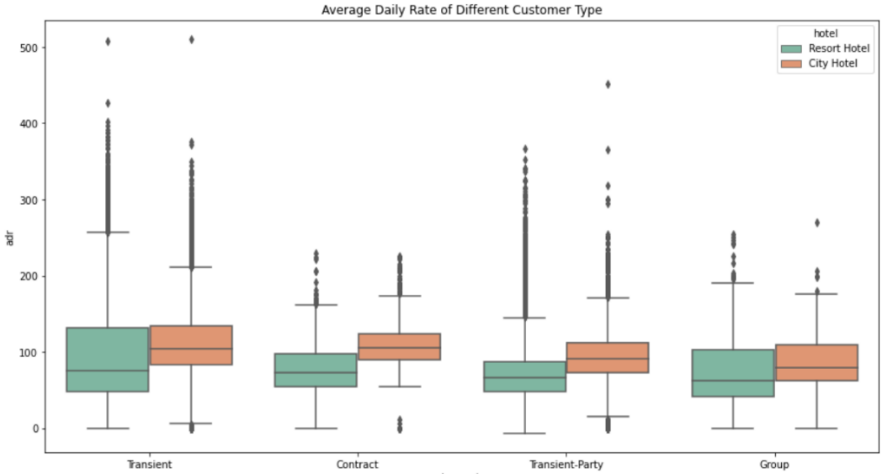
D.


City Hotel Customer Type
Transient
Transient-Party
Contract
Group
69.88%
26.96%
2.59%
0.57%
Resort Hotel Customer Type
71.85%
21.67%
5.59%
0.88%

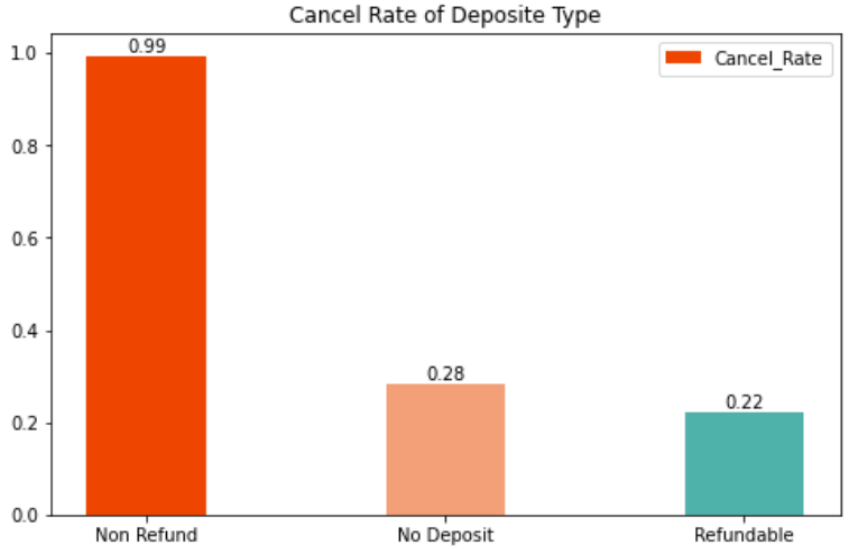
E.


City Hotel Market Segment
Online TA
Offline TA/TO
Direct
Groups
Corporate
Complementary
Aviation
52.47%
20.71%
10.90%
9.41%
5.07%
Resort Hotel Market Segment
39.67%
21.89%
19.47%
11.62%
6.77%
0.58%

F.


Average Daily Rate of Different Customer Type
hotel
Resort Hotel
City Hotel
adr
Transient
Contract
Transient-Party
Group

G.


Cancel Rate of Deposite Type
Cancel_Rate
0.99
0.28
0.22
Non Refund
No Deposit
Refundable

3. Insights:

A. The booking volume and cancellation rate of the City Hotel is much higher than that of the Resort Hotel. The hotel should conduct customer surveys to gain an in-depth understanding of the factors that cause customers to give up on bookings in order to reduce customer cancellation rates.

B. Hotels should make good use of the peak tourist season of July and August every year. They can increase prices appropriately while ensuring service quality to obtain more profits and conduct preferential activities during the off-season (winter), such as Christmas sales and New Year activities, to reduce the Hotel vacancy rate.

C. Hotels need to analyze customer profiles from major source countries such as Portugal and the United Kingdom, understand the attribute tags, preferences, and consumption characteristics of these customers, and launch exclusive services to reduce customer cancellation rates.

D. Since individual travelers are the main customer group of hotels and have high consumption levels, hotels can increase the promotion and marketing of independent travelers through online and offline travel agencies, thereby attracting more tourists of this type.

E. The cancellation rate of new customers is 24% higher than that of old customers. Therefore, hotels should focus on the booking and check-in experience of new customers, and provide more guidance and benefits to new customers, such as providing discounts to first-time customers and conducting research on new customers. Provide feedback on satisfaction and dissatisfaction with your stay to improve future services and maintain good old customers.

F. The cancellation rate of non-refundable deposits is as high as 99%. Hotels should optimize this method, such as returning 50% of the deposit or canceling this method directly to increase the occupancy rate.

G. The cancellation rate of room types A and G is much higher than that of other room types. The hotel should carefully confirm the room information with the customer when making a reservation, so that the customer can fully understand the room situation, avoid cognitive errors, and at the same time be able to understand the room facilities. Optimize and improve service levels.

H. Opportunity: The seasonal insight regarding peak and off-peak seasons can guide promotional strategies and dynamic pricing. With data on where the guests are coming from, the type of rooms they prefer, and the market segment they belong to, there's an opportunity to create targeted marketing strategies and personalized services, enhancing customer satisfaction and retention. Through the analysis of booking trends, cancellations, and customer preferences, there is an opportunity to create a predictive maintenance model that helps optimize hotel operations, inventory management, and staffing, thereby

saving costs. Analyzing features like ‘lead_time’, ‘previous_cancellations’, etc., can give deep insights into the likelihood of booking cancellations, helping in better capacity planning, and reducing the chances of revenue loss.

I: Challenges: Creating meaningful features from the existing dataset might be challenging. Identifying the right combinations or transformations that enhance the predictive power of the model would require deep understanding and experimentation. Given the diverse set of features, choosing the right model that can handle the complexity of the dataset and provide accurate predictions might be challenging. It would involve rigorous experimentation with different modeling techniques and parameter tuning.

4. Recommendation:

A. Handling Missing Values: Investigate the reason behind missing values in columns such as agent and company. Depending upon the reason, decide whether to impute them using statistical methods or remove the missing entries. For the ‘children’ and ‘babies’ columns, missing values can possibly be filled with '0', assuming the absence of an entry indicates no children or babies are accompanying.

B. Outlier Detection and Treatment: Conduct a boxplot analysis for numerical features like ‘lead_time’, and ‘adr’ to identify any outliers. Depending upon the analysis, I might decide to cap the outliers or remove them to prevent distortion in the model.

C. Feature Engineering: Create new features like ‘total_stay_days’ (sum of stays_in_weekend_nights and stays_in_week_nights) and ‘total_guests’ (sum of adults, children, and babies) to provide more detailed insights. Convert ‘arrival_date_year’, ‘arrival_date_month’, and ‘arrival_date_day_of_month’ into a single date column to facilitate time series analysis and identify trends more effectively.

D. Time Series Analysis: I will plan to conduct a time series analysis, ensure to set up the data in chronological order, and possibly create additional features like season or ‘week_of_the_year’ to identify patterns and trends over time.